

# **The Role of Dynamic-Network Multi-Agent Models of Socio-Political Systems in Policy<sup>1</sup>**

CASOS Technical Report

Marcus A. Louie and Kathleen M. Carley

January, 2007

CMU-ISRI-07-102

Carnegie Mellon University  
School of Computer Science  
ISRI - Institute for Software Research International  
Center for Computational Analysis of Social and Organizational Systems  
(CASOS)

## **Abstract**

Researchers and policy makers are increasingly turning to multi-agent and dynamic-network multi-agent models to study real-world systems. The models hold particular appeal because of their intuitive representation of complex real-world systems that can be thought of as “complex systems.” Both policy makers and those affected by the policies influenced by these models often question whether a model is valid. We explore the intended use of these models, the extent to which they can be validated, and the consequent implications for their use in setting policy. We ground the analysis using a dynamic-network multi-agent model we are helping to develop called the Regional Threat Evaluator (RTE), applied to data from Indonesia and Thailand. We find that there are three core difficulties in validating these models: defining the appropriate operating domain, data availability, and validating a model that integrates multiple theories.

---

<sup>1</sup> Acknowledgement: This work was supported in part by the AFOSR under a MURI with GMU (600322) on Adversarial modeling and the National Science Foundation IGERT in Computational Analysis of Social and Organizational Systems (NSF DGE-9972762). At the conceptual level, RTE is based, on an earlier model called Acumen, developed under DARPA. Additional support was provided by CASOS – the center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University (<http://www.ices.cmu.edu/casos>). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Air Force Office of Sponsored Research, the National Science Foundation, DARPA or the U.S. government.

Report Documentation Page			Form Approved OMB No. 0704-0188		
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE <b>JAN 2007</b>		2. REPORT TYPE		3. DATES COVERED <b>00-00-2007 to 00-00-2007</b>	
4. TITLE AND SUBTITLE <b>The Role of Dynamic-Network Multi-Agent Models of Socio-Political Systems in Policy</b>			5a. CONTRACT NUMBER		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER		
6. AUTHOR(S)			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>Carnegie Mellon University, Computer Science Department, Pittsburgh, PA, 15213</b>			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSOR/MONITOR'S ACRONYM(S)		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S)		
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>					
13. SUPPLEMENTARY NOTES <b>The original document contains color images.</b>					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES <b>41</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

**Keywords:** multi-agent modeling, validation, social network modeling, simulation, social systems

# Table of Contents

1	Introduction.....	Error! Bookmark not defined.
2	Characteristics of Multi-Agent and Dynamic-Network Multi-Agent Models.....	Error! Bookmark not defined.
	<b>Bookmark not defined.</b>	
3	Validation Continuum.....	Error! Bookmark not defined.
3.1	Types of Validation.....	Error! Bookmark not defined.
3.2	Types of Questions .....	Error! Bookmark not defined.
3.3	Synthesizing Validation Types and Questions .....	Error! Bookmark not defined.
4	The RTE: A Case Example.....	Error! Bookmark not defined.
5	Traditional Validation Techniques .....	Error! Bookmark not defined.
5.1	Operational Validation Techniques .....	Error! Bookmark not defined.
6	Validation Methods for DNMA Models: Challenges and Implications ....	Error! Bookmark not defined.
	<b>not defined.</b>	
6.1	Conceptual Validation .....	Error! Bookmark not defined.
6.2	Operational Validation.....	Error! Bookmark not defined.
6.2.1	Calibration.....	Error! Bookmark not defined.
6.2.2	Replication .....	Error! Bookmark not defined.
6.2.3	Hypothetical.....	Error! Bookmark not defined.
6.2.4	Forecasting.....	Error! Bookmark not defined.
6.3	Data Validation .....	Error! Bookmark not defined.
7	Validating an Integrated Model of Nation-State Failure: Overview of the Issues .....	Error! Bookmark not defined.
	<b>Bookmark not defined.</b>	
7.1	Defining the Operating Domain.....	Error! Bookmark not defined.
7.2	Applicable Techniques of Validation .....	Error! Bookmark not defined.
7.3	Theory Integration and Validation by Parts.....	Error! Bookmark not defined.
8	Conclusions.....	Error! Bookmark not defined.
	References.....	Error! Bookmark not defined.
	Appendix A: Micro-Theories in the RTE .....	Error! Bookmark not defined.
A.1	Deciding to Take Action.....	Error! Bookmark not defined.
A.2	Individual Action Selection .....	Error! Bookmark not defined.
A.3	Social Influence on Action Selection – Creation of the Final Attractiveness....	Error! Bookmark not defined.
	<b>Bookmark not defined.</b>	
A.4	Dynamics of Social Influence.....	Error! Bookmark not defined.
A.5	Weight and Adjustment .....	Error! Bookmark not defined.
A.6	Social Influence on Tension Adjustment.....	Error! Bookmark not defined.
	Appendix B: Model Parameters.....	Error! Bookmark not defined.
B.1	Entity Variables .....	Error! Bookmark not defined.
B.2	Provinces.....	Error! Bookmark not defined.
B.3	State.....	Error! Bookmark not defined.
	Appendix C: Parameter Space Constraints .....	Error! Bookmark not defined.

# 1 Introduction

Improvements in, and access to, computing power have increased the use of multi-agent systems in the policy arena. Multi-agent systems are touted as the method for analyzing “complex social systems,” particularly those characterized by multiple interacting parts and non-linear behavior. Multi-agent systems are appealing as real-world systems because they can be intuitively represented using the characteristics and behavior of individual agents as the basic building blocks. When relationships between agents and their dynamics are important in explaining real-world systems, dynamic-network multi-agent models hold particular appeal. As a result multi-agent models are being used in many diverse policy domains: e.g., civil violence (Epstein and Steinbruner et al. 2001), the spread of infectious disease (Carley and Fridsma et al. 2006), the disappearance of the Anazasi Indians (Dean 2000), the effects of government policies on the transportation of goods (Bergkvist 2004), and the effects of mutual influence on domestic water demand (Moss and Edmonds 2005).

Researchers and policy makers are turning to these models for reasons of ethics, cost, timeliness and appropriateness. In some systems, such as those modeling the spread of infectious disease, testing experimental conditions would put the safety of people at-risk, creating an ethical problem. In other cases, real-time evaluation of an existing system may be prohibitively long; whereas, simulation allows for rapid assessment. Simulation is also used when the cost of collecting data on the dependent variable is prohibitively expensive, or there are a large number of experimental conditions to test. For example, in a disaster, simulation is often the method of choice as there is often a need to rapidly evaluate many previously unexamined alternatives. In all these cases, since the real-world system under study is considered a complex, non-linear dynamic system, multi-agent simulations are often used as it is considered to have the appropriate level of complexity.

Decision-makers who use these models to inform policy and the people who are affected by decisions based on these models often question whether a model is valid. As the use of these models has become more prevalent a growing concern has arisen with how to validate such models. These large-scale multi-agent models represent a new approach to simulation for which traditional validation methods are not always applicable. From a history of science perspective it is important to note that the most advanced methods of validation were developed in engineering fields for assessing models of technical systems that followed fundamental physical laws. In contrast, these large-scale multi-agent systems are used for examining socio-political systems where the fundamental underlying laws do not exist or are at least unknown. We need to ask what is an appropriate validation process for such models?

Burton and others (Sargent 1992; Burton and Obel 1995; Thomsen and Levitt et al. 1999; Bigelow 2003; Burton 2003) have suggested that models and the appropriate level of validation are tied to the context in which these models are to be used. Thus, the intent of the users for the model should guide both the level of detail in the model and the extent to which it is validated. We ask what is the intended use of these large scale multi-agent models, the extent to which they should be validated, and how they should be validated. Further, what implications does the achievable level of validation have for the role these models should play in influencing policy in various domains?

This paper is organized as follows. We first review key characteristics of multi-agent and dynamic-network multi-agent models and then discuss how the purpose of a model influences what are appropriate types and levels of model validation. Then we describe a dynamic-network

multi-agent model, the Regional Threat Evaluator (RTE), that we have developed, coupled with datasets on Indonesia and Thailand which serve as an example for the remaining analysis. We then examine different approaches and techniques used for validating systems and ask what are the problems in applying these approaches to multi-agent simulations. We illustrate these problems using RTE and the associated data. Note, this paper is not a presentation and validation of the RTE model. Rather, this paper is an analysis of the appropriateness of standard validation procedures for large scale multi-agent models such as the RTE.

## 2 Characteristics of Multi-Agent and Dynamic-Network Multi-Agent Models

We often observe a real-world phenomenon and then hypothesize about causes. Many computational models investigate problems from the other direction by first defining the rules, the process, or the behavior, and then observing the phenomenon<sup>2</sup>. Multi-agent models use individual agents and their behaviors as the basic building blocks for this investigation. In these models, individual agents are constructed as a set of rules that define behavior, then let go to operate in parallel. Group phenomena emerge from the interactions of the individual agents.

In traditional multi-agent models agents exist on a grid and have individual rules that guide their behavior. Typical examples are cellular automata models such as Sugarscape (Epstein and Axtell 1996) and the Axelrod Cultural Model (Axelrod 1995). To the extent there are networks connecting agents they are static and grid based. In contrast, in dynamic-network multi-agent models, agents exist in multi-dimensional socio-political-technical space where the networks among agents co-evolve with the agents. This is important for systems where the shifts in the relationships between the agents are important in explaining the system's behavior. An example of such a system is BioWar (Carley et al, 2005) where the disease network evolves as agents change patterns of interaction as they get sick.

Multi-agent models and the related dynamic-network multi-agent models have several key characteristics:

- *Bottom-up approach to theory development.* The behavior of the system is derived from the individual behaviors of the agents. Theory for how the agents should behave is drawn from extant literature or is hypothesized.
- *Capable of expressing non-linear behavior.* Unlike in linear systems where a change in one factor causes a proportional effect, changes in a factor used in a multi-agent model may result in a very large effect, no effect, or a proportional effect.
- *Path-dependent.* While not a requirement, in multi-agent models, states are often dependent on the previous states of the model; history matters.
- *Boundaries are subjective.* Descriptions of multi-agent models usually describe what the agents are, their network topology, the variables that describe the agents, the invariants of

---

<sup>2</sup> Lomi and Larson describe the two approaches as the forward and backward problems. For multi-agent systems, the forward problem is “given a set of assumptions about individual decision rules and problem solving procedures, can we determine (predict) the aggregate properties of a system generated by the repeated interaction among individual units adopting these rules and procedures?” The backward problem is “given the observable regularities in the behavior of a composite system (e.g. an organizational field or a market) can we specify a set of rules or procedures that – if adopted by all elementary units – induce and sustain these regularities?”

the variables, and the individual behaviors or rules governing the agents. Beyond that, specification of where the model ought to be used is subjective.

- *High-dimensionality.* Both multi-agent and dynamic-network multi-agent models have the potential for a large input space, but the potential is greater in the case of dynamic-network multi-agent models where the pair-wise relationships among agents in a network need to be described. The high-dimensionality has implications for over-fitting the data.

Being principally concerned about the study of socio-political systems, we compare the multi-agent models to other classes of models that are used to studying such systems. System dynamics models (Forrester 1961) have been used to model a number of socio-cultural systems including how stocks of commercial structures, housing structures, and a population in an urban environment change over time accounting for various population flows and municipal policies (Forrester 1969), epidemics of infection disease (Kermack and McKendrick 1927), and a number of business related applications (Forrester 1971). Specifying a system dynamics model starts by describing a high-level structure for how the system behaves, adding more detail to each component iteratively. It requires that those constructing the model know or have hypothesized about how the system behaves in the real-world. This top-bottom approach is in contrast to the bottom-top design of multi-agent models where the focus is on specifying the behavior of the individual parts, or agents. The agents are then linked together either on a grid or in a network topology to form the larger system. Game theory is used as a way of modeling strategic behavior between two rational actors seeking to maximize payoffs, but derivatives have been developed to accommodate multiple actors as well.

Table 1 summarizes the differences between these types of models. Key differences are bottom up approach to theory, subjective boundaries, high dimensionality and strong path dependence. These differences suggest that validation issues will be somewhat different for multi-agent than other types of complex system models.

Attribute	DNMA	System Dynamics	Regression	Game Theory	Multi-agent
Multi-theoretical	Yes	Yes	N/A	No	Yes
Multi-level	Yes	Yes	Yes	No	Yes
Emergent behavior	Yes	Yes	No	No	Yes
Bottom up approach to theory development	Yes	No	N/A	No	Yes
High dimensionality	Yes	Sometimes	No	No	Sometimes
Agents exist in a dynamic network	Yes	N/A	N/A	No	N/A
Capable of expressing non-linear behavior	Yes	Yes	No	No	Yes
Enable analysis of hypothetical situations	Yes	Yes	Limited	Limited	Yes
Path-dependent	Yes	No	No	No	Yes
Predict behavior people	Yes	Yes	Yes	Yes	Yes
Boundaries are subjective	Yes	No	No	No	Yes
Predict behavior of technology	Yes	Yes	Yes	No	Yes
Enables dynamic analysis	Yes	Yes	No	No	Yes

**Table 1. Comparison of model classes**

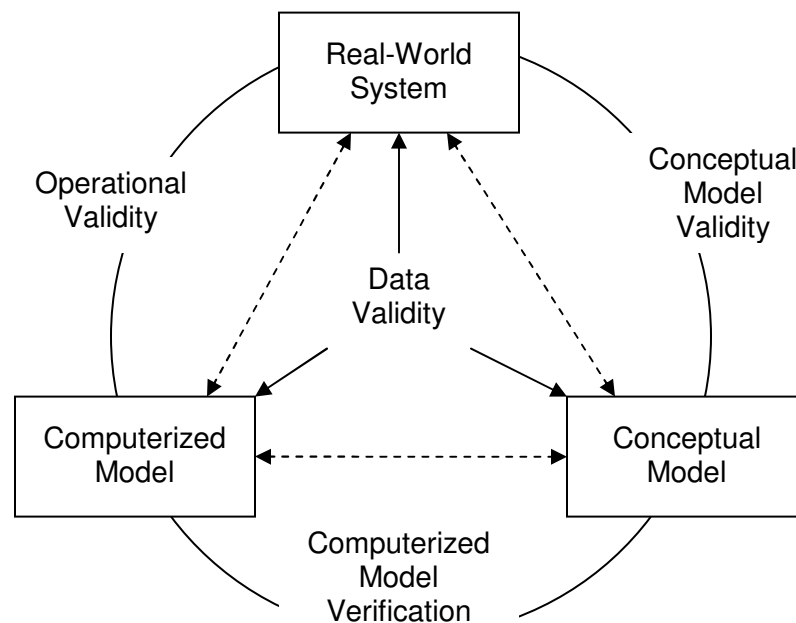
### 3 Validation Continuum

The validation process should be tied to the purpose and the context for which the model is being developed (Sargent 1992; Burton and Obel 1995; Thomsen and Levitt et al. 1999; Bigelow 2003; Burton 2003). Sargent provides an overview of different validation techniques, each providing different types and levels of validity (Sargent 1992). He notes that the desired



level of validity is determined on the purpose of the model, but does not attempt to describe in detail what different purposes are and how they relate to the validation process. Burton complements Sargent's work by describing types of questions that are asked of simulation models (Burton and Obel 1995; Burton 2003) while recognizing that the level of validation is still dependent on the question, or purpose, of the model.

In this section, we synthesize this related work and organize the types of questions that are asked of models and associate them with the types of validation that are appropriate for each type of question. To assist in giving structure to the synthesis, we use a conceptual description of the simulation-model development and validation process given by Banks, Gerstein, and Searles (Banks and Gerstein et al. 1987). Figure 1 is a representation of the conceptual components and steps of the process, recreated from Sargent (1992).



**Figure 1. Conceptual depiction of the components of model development and validation. Figure is recreated from Sargent. (1992).**

### 3.1 Types of Validation

Figure 1 shows three different parts of a simulation model that can be validated. This section identifies the different types of validation that can be performed for each of these model parts. Section 5 is devoted to typical techniques used to establish different types of validity. Conceptual validity is determining the extent to which the model theories and the underlying assumptions are appropriate for the purpose of the model. Determining the validity of data involves making sure that the data are appropriate for the purpose of the model, that a sufficient amount of data exists to build and validate the model, and that the data are accurate with respect to the real system. The operational validity of the model is determining the extent to which the model produces output that matches the real system under investigation for the purpose in which it was developed.

Before a dynamic-network multi-agent model is written, a conceptual model is usually produced which identifies the assumptions, relevant entities, their actions, and relationships among entities. The next step is to formalize the concepts into mathematical relationships. Validating the algorithms answers questions relating to whether the equations and computational procedure used represent the conceptual model. Validation of the conceptual model is done by subject matter experts.

Though this paper is focused on validating dynamic-network multi-agent models, we mention data validity because the quality of available data often constrains both model development and operational validation. In simulation models where the data do not come directly from the real-world, the input data are subject to strong biases that reflect the background of the data collectors. It is necessary to minimize the variation of model output produced by different biases of data collectors. Ideally, a formal data collection method should be established.

Typically, in a simulation model one attempts to match the real system being investigated in a sequence of steps. Often the first step is to calibrate the model to show the extent to which it can reproduce some target system. To determine whether the model has been overfit to that target system, usually one or more target systems are used to determine how well the model generalizes to other systems. Historical data from the target systems is usually used to calibrate the model and to see how well it generalizes. Prospective data from the system can also be used, and using this type of data from the system is referred to as forecasting. Thus, the goal is see how well the model can predict future conditions of the target system. A related form of forecasting attempts to not just forecast the future, but to attempt to change the future based on results from the model. In forecasting with an intervention, the intervention is used in the real system. The level of validity that has been achieved in previous tests will determine whether real resources and people will be committed to implementing the intervention.

### 3.2 Types of Questions

We commonly think of two broad categories of questions in policy: the positive and the normative. In the positive, we observe the world, create a description and then seek an explanation. Theoretical models are proposed and then tested by creating hypotheses to answer questions we do not know. In the normative, we seek to describe what is good, what should be done. This requires a confirmation that not only does the model make sense in the realm of the positive, but that its boundaries of explanation overlap to the subset of the normative for which the model is being applied. Somewhere between these questions is a third category of questions, questions of what is plausible (Burton 2003). The plausible are explorations of *what might be*.

Simulation models are particularly useful in this context. We include system dynamics, multi-agent, and dynamic network models as types of simulation models. In the real-world we can devise controlled laboratory experiments if we would like experimental control over explanatory variables, but in much larger contexts such as processes of state-failure or disease epidemics, we cannot easily manipulate the environment to conduct a controlled experiment and if we could, it is not always appropriate or ethical as mentioned previously. Further, we might think about data coming from the real-world as a single data stream. With the exception of comparative case studies and extrapolation from data, we are not able to explore the plausible beyond what has already happened. Asking plausible types of questions allow the use of theory and simulation models to look beyond what has already happened to ask what could happen.

### 3.3 Synthesizing Validation Types and Questions

In Table 2, we organize the types of questions along with the types of validation to produce a concept of how the type of question a model is designed to answer influences the types of validation that the model should go through. Certainly, the model is being designed to help inform policy choices, but it is not being designed to explicitly ask “what is good,” or normative types of questions. A judgment of what is good is reserved for the policy makers.

		TYPE OF QUESTION		
		Positive	Plausible	Normative
TYPE OF VALIDATION	<b>Conceptual</b>	Representative	Yes	Yes
		Proper Algorithms	Yes	Yes
	<b>Operational</b>	Calibration	Yes	Yes
		Replication	Yes	Yes
		Forecasts	Yes	Yes
		Hypothetical	No	Yes
	<b>Data</b>	Intervention	No	Yes
		Sensitivity to Parameter Bias	Yes	Yes

**Table 2. Types of questions and validation steps that are appropriate for each question. Table cells marked "Yes" indicate that the validation is appropriate for increasing the degree of validity for the corresponding type of question.**

The RTE is being developed to answer plausible questions such as: How is the likelihood of state failure change in the face of a major natural disaster? How do increased levels of terrorism in the country affect the likelihood of state failure? The model is not intended to answer normative questions. In a model of state failure, there are a number of indicators that people pay attention to, but the values of those indicators that are “desirable” may vary between decision makers and may further be constrained by the context.

## 4 The RTE: A Case Example

The RTE Model is based on the integration of multiple theories of social, psychological and economic behavior that collectively account for why an agent takes action and what action gets taken. It is currently being developed from a similar model based on urban threat environments. Both models were developed as special-purpose models to answer specific types of questions.

The basic idea is that inter-group conflict is due to a combination of tension (Horowitz 1985) and social comparison (Festinger 1954), the effects of which can be modulated by social pressure (Friedkin 1998). Agents who are more tense and who see themselves at more of a disadvantage relative to others are more likely to engage in hostile actions; whereas, lower tension and higher advantage lead to non-hostile actions. Agents who have influence over the agent in question can use that influence to escalate or de-escalate the impact of tension and social comparison. Specifically, an agent who is influenced by others who themselves are tense or feel deprived will feel more tense and deprived than will an agent surrounded by others who are less tense or less deprived. Social influence derives from shared attributes such as culture,

knowledge, borders, and goals and co-evolves with those attributes (Carley 1991). It follows that the more heterogeneous a population and the more the lines of differentiation line up the greater the potential for hostility (Blau 1977). When agents decide to take action, the action and target are selected using a bounded-rationality cost-benefit analysis (Mishan 1973; Simon 1982; Allison and Zelikow 1999) subject to resource constraints (Pfeffer 1978). The costs and benefits of taking a particular action against a particular target are also modulated by social influence. Thus, agents are more likely to take the kinds of actions against the kinds of targets that social pressure suggests are appropriate and will be sanctioned by other agents for inappropriate action or target choice.

The RTE Model is a multi-agent network model of state failure (see Figure 2). In this model, boundedly rational agents interact and take actions to achieve goals. When agents act they take into account what resources they have available, the cost and benefits of the action, and the opinions of others whom they are influenced by. These actions influence the likelihood of state failure. State failure is measured using nine factors and a composite indicator. These factors are lack of state legitimacy, potential for province secession, hostility, tension, level of corruption, level of terrorist activity, level of criminal activity, level of foreign military aid, and lack of essential services. State failure is also measured at the province level using similar indicators.

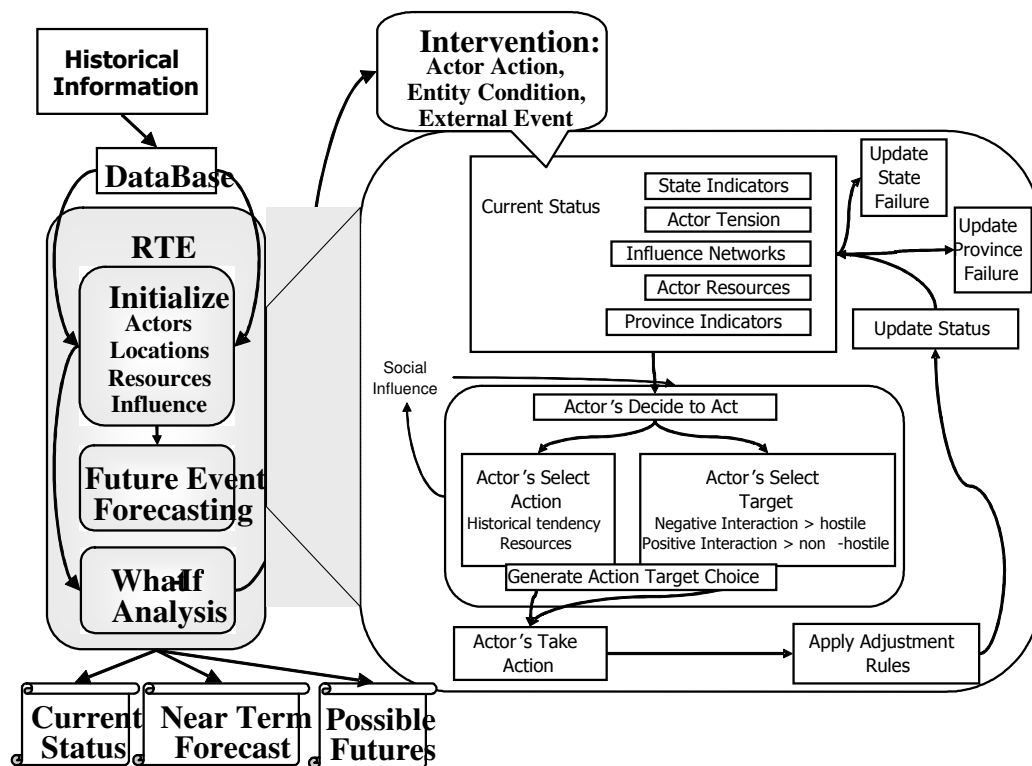


Figure 2. Top-level view of the Regional Threat Evaluator (RTE)

The model is initialized using real world data and then the agents proceed to interact and take actions which consume or generate resources. Activity at the agent level then leads to changes in these agents, their resources, the non-agent targets, and indicators of state stability. For example, forced migration of a population from one province to another is likely to decrease

tension in the province left, increase tension and hostility and decrease essential services in the province migrated to, increase tension in the population that migrated and decrease their resources.

Data is used to set the initial conditions of the model, inform the scenario, and do limited validation came from a 32 different sources. These sources included: 6 national agencies (e.g., Indonesia-Tourism.com), 6 NGO's (e.g., United Nations and World Bank), 4 US Agencies (e.g., CIA and Dept. of Energy), 6 News Services (e.g., Bangkok Post and BBC), 10 Research and Academic Institutions (e.g., Terrorism Knowledge Base and Institute of Southeast Asian Studies), and 2 Corporate/Labor groups (Netcraft and International Telecommunication Union). In addition, specific information on the relevant entities and provinces came from various on-line news (e.g., Washington Post) and web-services (e.g., Wikipedia) from both US and foreign media. Illustrative websites used include <http://www.uis.unesco.org>, <http://www.tkb.org>, [http://www.undp.or.id/pubs/ihdr2001/ihdr2001\\_full.pdf](http://www.undp.or.id/pubs/ihdr2001/ihdr2001_full.pdf), and <http://www.unescap.org/esid/psis/population/database/thailanddata/thailandfacts.htm>. This data were used as a basis for 150 state indicators, 60 province/region indicators, and 30 entity indicators used to initialize the simulation model.

## 5 Traditional Validation Techniques

### 5.1 Operational Validation Techniques

Law and Kelton describe several commonly techniques to perform operational validity (Law and Kelton 1999). To this list we add model-to-model comparison (Axtell 1995; Burton 2003). Briefly, we describe each of these techniques and discuss how the purpose of the model and availability of data affects the ability of certain types of techniques.

*Inspection.* The model results and variable relationships, including input-output relationships, are compared against the model developer's expectations. The expectations are based on the conceptual model that was developed before the implemented computational model.

*Face Validation.* In validating the operational abilities of the model, subject matter experts are used to judge how well the model compares to the real system. Face validation can be used to compare the model to known system behavior or to prospective system behavior with or without an intervention. Face validation results in one or more experts each describing the extent to which the model outputs are as expected.

*Confidence Interval.* When a large amount of data exist from the model and from the real-world system then a confidence interval can be constructed that compares the independent sets of data from the model to independent sets of data from the real-world system. The method provides a magnitude of how far off the model is from approximating the real-world system.

*Time Series.* In time-series approaches, a set of output data from the model and the real-world system are compared to each other to see how well the two system's behavior agree with each other time.

*Point Comparison.* This comparison has a stringent condition that the model is deterministic. When this is the case and when data exist to compare model output to, then a comparison can be made to see how different the model predictions are from the real data and under what conditions.

Which real-world data are available affects the type of validation that can be done. Kleijnen enumerates three conditions of data availability and proposes statistical techniques that can be used depending on the situation (Kleijnen 1999): no (or very limited) real-world data,

data on the real-world output only, and data on both the real-world output and the real-world input. In the case of the RTE, though output data from the real-world exist in the form of published reports and public news sources, the data are qualitative and not amenable to statistical techniques. Statistical techniques that do not require output data from the real-world can still be performed and can still provide an increased level of confidence that model is behaving as expected. For example sensitivity analysis can be used to check to see if factors have effects that agree with qualitative descriptions in the literature or with subject matter experts' knowledge.

Table 3 summarizes the types of data that are available to validate the RTE for each of the types of statistical techniques discussed in Law and Kelton.

<b>Technique</b>	<b>Indonesia</b>		<b>Thailand</b>	
	<b>Type/Quality of Available Data</b>	<b>Nature of Data</b>	<b>Type/Quality of Available Data</b>	<b>Nature of Data</b>
Face Validation	SME (Military personnel)	Human judgment	SME (Thai specialist Anthropologist)	Experts disagree
Inspection	Public news sources, published reports	Human judgment, incomplete	Public news sources, published reports	Human judgment,
Confidence Interval	N/A	N/A	N/A	N/A
Time Series	Public news sources	Possible exogenous factors not included	Public news sources	Possible exogenous factors not included
Point Comparison	N/A	N/A	N/A	N/A
Model-to-Model	High Level Results	Results generated by 3 other models	High Level Results	Results generated by 3 other models

**Table 3. Summary of the types of data available for validating the aggregate behavior of the model.**

## **6 Validation Methods for DNMA Models: Challenges and Implications**

Models can go through several broad evaluation processes at different times during development. Operational validation is conducted in the form of calibration, replication, forecasting, hypothetical analysis, and forecasting with an intervention. Conceptual validation is performed as the model is developed and new dynamics are added. The degree to which each of

these is achieved impacts not only how credible the model is to others, but consequently how useful it is in a policy-context role. In the case of the RTE, each of these categories of evaluation suffers in some way.

Examples of each evaluation process are given to demonstrate what is achievable and what the limitations are.

## 6.1 Conceptual Validation

Development and validation of the RTE has progressed in parts in an attempt to narrow the space of parameters that need to be calibrated at any one time. For example, when constructing rules that affect corruption levels, we turned to regional experts and published reports on Indonesia to hypothesize about the relevant factors driving corruption levels. While a lack of quantitative data on corruption levels prevents us from comparing model output to data from the real world using statistical analyses, sensitivity analysis can support validation by showing whether factors have effects that they are expected to have. Based on input from the regional experts and reports, the following variables and equation is used as a first-order approximation to update levels of corruption (Equation 1):

- Initial corruption level,  $C_0$
- Adjustment to corruption due to aid type and level,  $A$
- Minimum level of acceptable service in the region,  $S_T$
- Current level of service in the region,  $S$
- Region volatility,  $V$

$$C' = C_0 + (1 - C_0) \cdot (S_T - S) \cdot A \cdot V \cdot C_0$$

**Equation 1. A modified version of the weight and adjustment formula is used to update most other variables in the model. This equation considers the difference between current levels of essential services versus what the region's perceived minimum level of essential services is. All variables are assumed to have values between 0 and 1 inclusive.**

Updates to corruption use a modified version of a weight and adjustment formula that is used to update most other variables in the model (Appendix A.5 Weight and Adjustment). The formula is modified to include the  $(S_T - S)$  term, which weights the adjustment by the difference in the minimum level of acceptable service in a region and the current level of services being provided. The modification is motivated by the idea that less aid is grafted when there is more of a perceived need for aid.

Figure 3 shows how the initial corruption level before the tsunami affects the future level of corruption once aid starts to flow into the state. The model was run using parameters for Indonesia. All tested variations in initial corruption levels show that future increases in levels of corruption start to happen around week 16. The stability of this feature in model output gives us confidence that initial corruption levels will not significantly alter when increased levels of corruption begin to occur. Regional experts agreed that changes in levels of corruption were low during the first 5 months as most people were in need of food, water, and medical care. Once reconstruction of housing and infrastructure begins, corruption becomes most prevalent as contractors are hired (Stansbury 2005). The government of Indonesia focused on immediate relief (food, water, medical aid, and temporary shelters) until June 2005, thereafter the focus shifted to rehabilitation and reconstruction. Thus, there is agreement in the literature on corruption, what is known about Indonesia's recovery plan, and the regional experts' beliefs

about corruption levels in Indonesia. The model's results approximated the expected rise in corruption levels, but predicted the rise about one month early.

Initial corruption levels did affect how quickly future corruption levels increased. It is expected that less corrupt countries would lead to less of an increase in future levels of corruption, but Figure 3 shows that lower initial levels of corruption lead to greater increases in corruption than greater initial levels of corruption do. The discrepancy indicates that the underlying conceptual model may be flawed. As a result the operating conditions should be constrained to countries with relatively high corruption levels (as in Indonesia and Thailand) or the theory should be amended to include countries with lower levels of corruption.

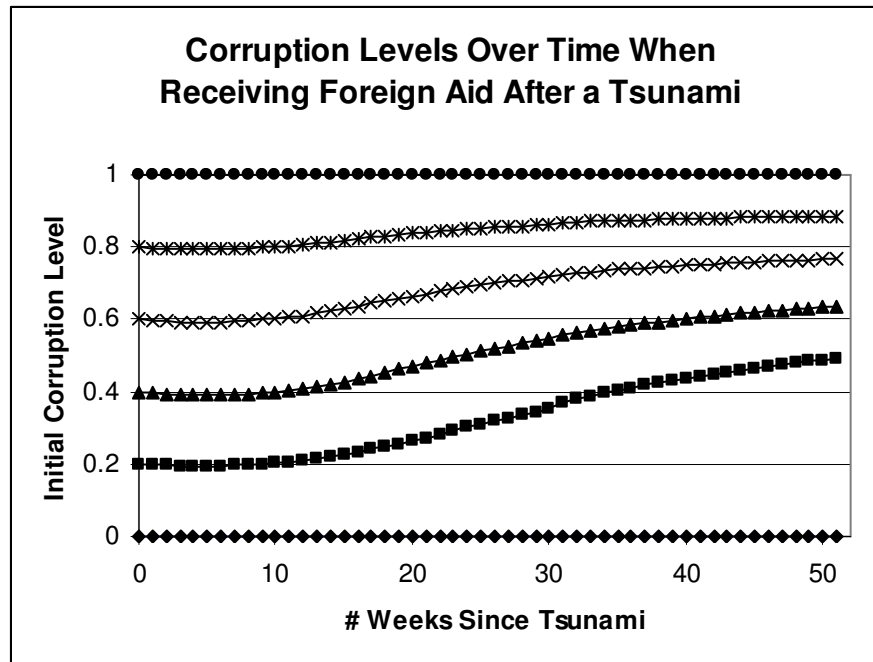


Figure 3. Initial corruption levels affect how quickly future corruption levels change over time. The rate of increase in corruption tapers off as time progresses.

Even when data are sparse and insufficient for doing quantitative analyses, conceptual validation can still be useful. For example, it may be sufficient to know that a change in an input parameter in a specific range can have a non-linear effect on one or more output measures. Further, because the conceptual model validation is derived from the mental models of the developers and those providing expertise, unexpected results challenge assumptions and either force a revision of the conceptual model or a deeper analysis at explaining the outcome.

## 6.2 Operational Validation

### 6.2.1 Calibration

Models of complex social systems tend to have very large parameter spaces, and the challenge becomes navigating the space with discipline and rigor. Optimization methods may help, but without constraints calibrated parameters may be logically inconsistent with respect to the system being studied. Yahja and Carley (Yahja and Carley 2006) explain an automated approach under development in which they combine optimization methods and knowledge



representation to explore the parameter space under constraints by a rule language expressed outside of the model. It is capable of automatically tuning the model and model rules to fit a data set. While this method was not used in the RTE, in Appendix C, we describe how we attempted to reduce and quantify the size of the space.

To calibrate the data, model output were compared to qualitative accounts of Indonesia before and the months following the tsunami on December 26, 2004. The tsunami condition was created by having the tsunami affect the 10 provinces most impacted by the disaster. In the model, the tsunami significantly reduced resources in the affected province and of the agents who live in those provinces. Ten essential services were significantly degraded. Aid flowing into the country was modeled after the stages of recovery, rehabilitation, and reconstruction that the Indonesian government set forth (Action Contre La Faim 2005). Indonesia is perceived to be one of the most corrupt governments<sup>3</sup>, and as such there was concern that a significant amount of aid flowing into the country would be siphoned off by different levels of the government and by the military (Batha 2005). Within the first five months after the tsunami, reports by aid-agencies have witnessed only small-scale corruption. There is concern, however, that rehabilitation and reconstruction projects which are long-term and expensive will be at more of a risk of having funds grafted (Simanjuntak 2005). One possible scenario is that aid funds are increasingly siphoned off after the initial months when corruption is low. Another possible scenario is that corruption levels have increased as aid has come in. This scenario would be supported by the idea that the reason there is little evidence for corruption is that aid agencies have an interest in not exposing corruption because doing so discourages donations. The output from the model is line with the former scenario and was assumed to tentatively match the real world.

Calibration also requires that there is a method to compare model output with data from the real-world system. Models of social systems often have variables that are not observable in the real-world so comparing model values to one or more proxy variables to see how well the model fits the data becomes an exercise of interpretation.

The RTE was calibrated using inspection and face validation with regional experts in Indonesia. Using inspection we compared the nine high-level indicator variables to the conditions that are described in both public news sources and published policy reports. The goal of the inspection was to approximate the conditions that were being reported in each of the two countries. Face validation of the model occurred after the model was inspected and tuned. In this stage, a description of the model and model results from the tsunami and aid scenarios were presented to a group of regional experts on Indonesia and Thailand. The members of the group were put together by the Defense Advanced Research Projects Agency. Representative affiliations include the Defense Intelligence Agency, Office of Naval Research, and the U.S. Pacific Command. Face validation by the regional experts focused on whether the conditions represented in the model produced the expected outcomes in terms of the indicator variables overtime.

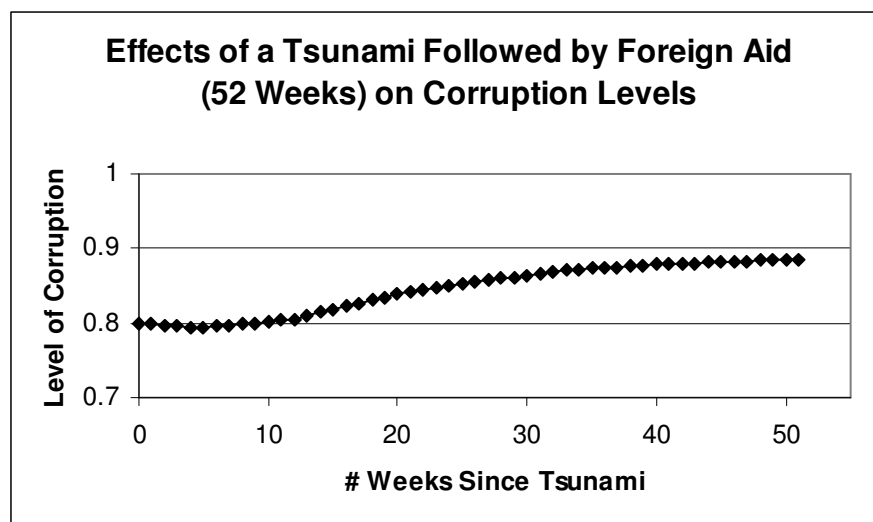
While inspecting the model, the following are the main effects we were trying to demonstrate in the case of Indonesia facing a tsunami:

---

<sup>3</sup> Transparency International constructs an annual ranking of perceived levels of corruption of governments. Scores relate to perceptions of the degree of corruption as seen by business people and country analysts and ranges between 10 (highly clean) and 0 (highly corrupt). Indonesia received a score of 2.0 in 2004.

- Foreign aid has a relatively small impact on levels of corruption<sup>4</sup> during the first few months after the tsunami, after which continued foreign aid will lead to higher levels of corruption.
- Hostility between the Free Aceh Movement (GAM) and the Indonesian military (TNI) will decrease.
- Overall likelihood of state-failure in the months following the tsunami will be slightly greater than what it was before the tsunami.

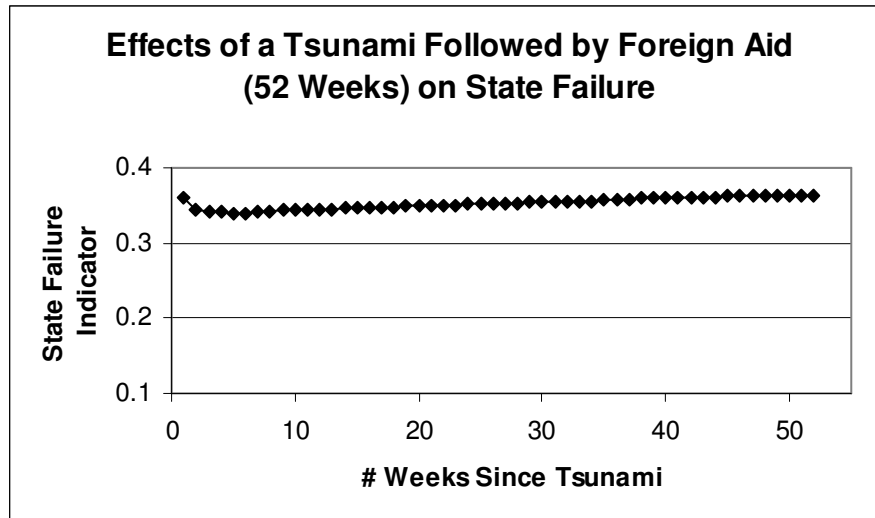
Figure 4 allows us to examine whether the model we developed and tuned matched these expectations. To examine the effects of foreign aid on levels of corruption, we plotted the level of corruption overtime for one year after the tsunami hit. Each time period in the model is calibrated to be approximate one week in the real world.



**Figure 4. The level of corruption increases in the first few weeks of foreign aid following a tsunami.**

To understand the effects that the tsunami had on the state failure indicator for Indonesia, we compared the value of the state-failure indicator before the tsunami and then the overtime value of the state-failure indicator overtime after the tsunami. Figure 5 shows that the value of the state failure indicator dips immediately after the tsunami and then recovers to approximately the same value within a year. The dip is due to an overall reduction in hostile activities across the state. While the direction of state-failure indicator was incorrect, the magnitude of change in the state-failure indicator was marginal. Nonetheless, an in-depth trace back of the process can help reveal reasons for the prediction.

<sup>4</sup> Transparency International defines corruption “as the misuse of entrusted power for private gain.” <http://www.transparency.org/faqs/faq-corruption.html>. Transparency International is a non-profit organization devoted to promoting governance free of corruption.



**Figure 5. The state failure indicator for Indonesia rises slightly when foreign aid comes in after a tsunami.**

Reports about the hostilities between GAM and TNI were shown to taper off after the tsunami (McCulloh 2005) and peace talks were started with GAM leadership, eventually reaching a memorandum of understanding in August 2005. The results in Figure 6 show the average number of actions the RTE predicts the TNI will take against GAM per week in two different scenarios. One scenario assumes no tsunami has occurred and can be interpreted as behavior before the tsunami. The other scenario interjects a tsunami. For the RTE to match what is expected, Figure 6 would need to show that the average number of actions the TNI took against GAM before the tsunami was greater than when the tsunami occurred. Figure 7 shows the corresponding graph for the number of actions GAM takes against TNI. A comparison between Figure 6 and Figure 7 shows that the RTE predicts GAM to be more active toward TNI than TNI was toward GAM, which is questionable given the known disparity in resources. Further, GAM was reported to have retreated in order to preserve resources.

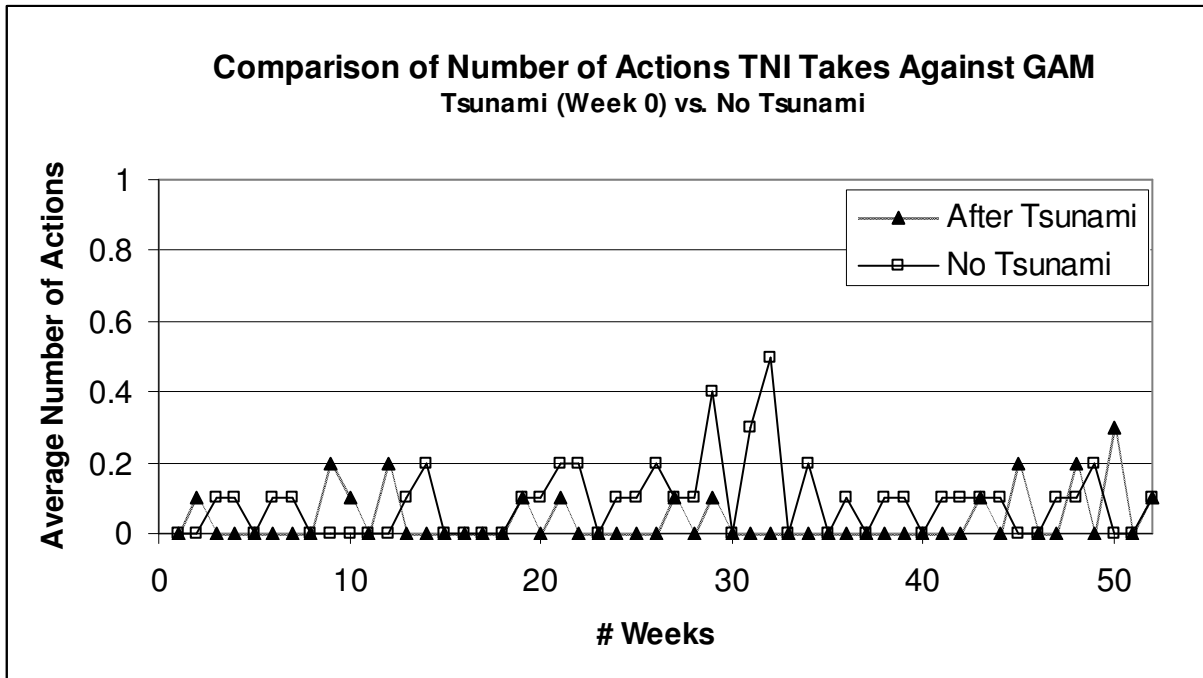


Figure 6. The RTE predicts that there is no difference in how often TNI takes actions against GAM either before or often an tsunami.

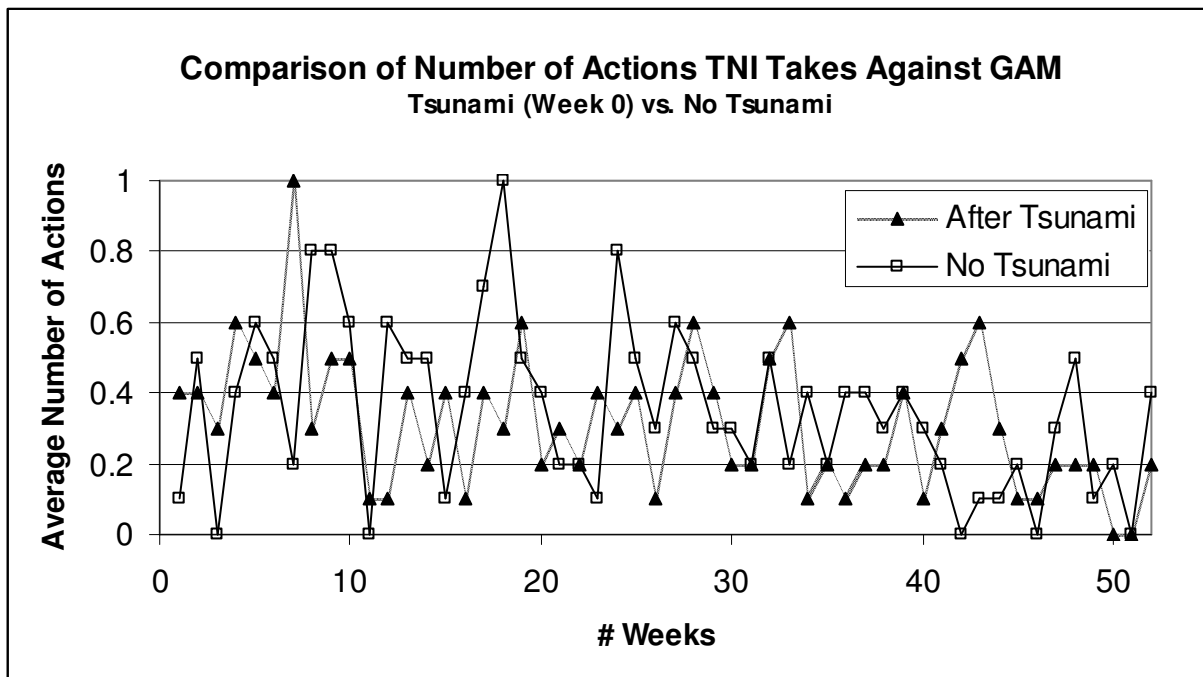


Figure 7. The RTE predicts there is no difference in how often GAM takes actions against TNI either before or often an tsunami.

The regional experts confirmed the main effects of the tsunami that the model generated. In addition to the main effects, the regional experts also reviewed the individual-behavior of the agents. The agents in the model are responsible for the dynamics of the model, so their behavior

was seen as crucial to understand whether the correct processes were being modeled. The agents' choice model and the structural-influence model were both presented (Appendix A). Some regional experts questioned the concept of action attractiveness being based off of a deterministic calculation of perceived impacts and perceived costs as being too closely related to a rational choice theory, which they believed was incorrect for this domain. These same people did agree that the components that made up of the attractiveness calculation captured the right factors. Everyone agreed that how agents are embedded structurally in the system has an important role in determining outcomes. Further, the network structure is recognized to be important at explaining dynamics at varying levels of conflict from urban operations (Medby and Glenn 2002) to intra-state and inter-state conflicts. Use of Friedkin's structural influence theory to capture the impact of network structure on choice, likelihood of acting, and affinity toward others was seen as appropriate.

It is useful to compare the tuning process of the RTE to a more traditional approach to understanding state-failure. Regression models have been employed by the State Failure Task Force (Goldstone and Gurr et al. 2000; Johnson 2004) to discover measurable characteristics of countries which affect the risk of state failure. Tuning their regression models amounted to creating a dataset of countries, their hypothesized relevant characteristics, and then using a partial set of the data to parameterize the model coefficients. The Task Force uses logistic regression models to calculate the odds of different events related state-failure. Thus, the tuning of the model parameters had a systematic method, namely, numerical analysis to iteratively select parameters that maximize conditional likelihood of state-failure given their hypothesized independent variables.

While it is a non-trivial task to calibrate a multi-agent model failing to adequately do so does not mean the model is without value. For example, sometimes a model cannot be calibrated to a real-world context because there is a lack of data. In the case of the RTE, we relied on subject matter experts to judge whether the model was able to appropriately match the real world. Even supposing none of the subject matter experts agreed that the RTE produced output that related to the real world, the model was still useful to sparking discussion of what the. While all models are likely to spark discussions of appropriate representation, data, and dynamics, the discussion is influenced by the type of model used. Use of multi-agent models are likely to stir up discussions of agent representation, relevant relationships, and the behavior of agents. These are discussions that are likely not to occur if a regression model were used. In short, use of a multi-agent model encourages discussions that are not likely to happen when other methods are used.

### 6.2.2 Replication

Once a model is tuned we would like to know how generally the model can be applied to answer the same questions in different contexts. For example, the BioWar (Carley and Fridsma et al. 2006) model was calibrated using Pittsburgh as a test scenario. Replication of other cities tested the model's underlying theories to see how generally they can be applied. Similarly, data from Thailand was collected to see how closely the model was capable of producing results that matched what is known about Thailand.

As with Indonesia, we examine the model's predictions of a tsunami on Thailand focusing on trends of corruption in Thailand and the composite indicator of state failure. Focusing on these two effects allows comparison of how well the model predicts similar phenomena in different contexts, in this case, between Indonesia and Thailand. To the extent

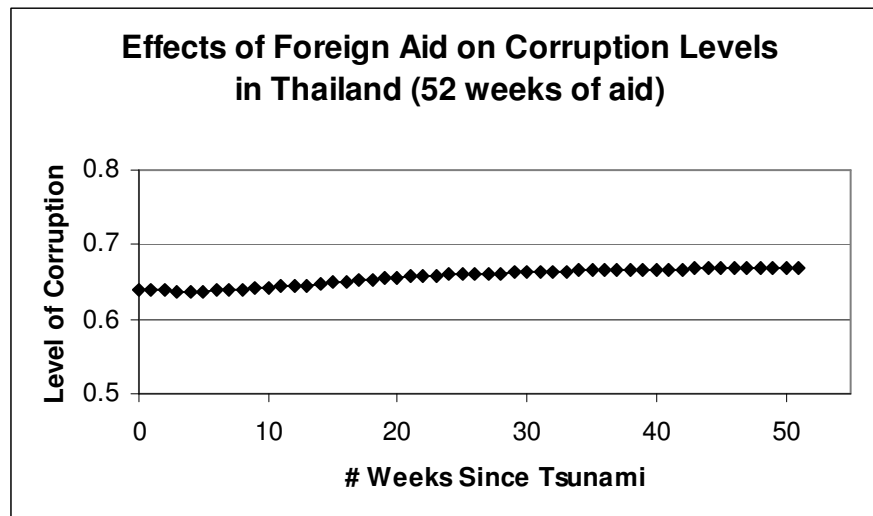
that the model is able to approximate the phenomena in both countries, the more we might believe that model's assumptions, representation, and dynamics are correct.

The following main effects of a tsunami are expected in Thailand.

- Foreign aid has a relatively small impact on levels of corruption. It is expected that corruption will be less in Thailand than in Indonesia.
- Overall likelihood of state-failure in the months following the tsunami will be slightly greater than what it was before the tsunami.

Corruption is not expected to be as great in Thailand as it is in Indonesia. Part of this is due to the perception that Thailand is less corrupt than Indonesia<sup>5</sup>, but also important is that Thailand has received much less aid due to the tsunami than Indonesia has<sup>6</sup>. Indications of state failure in Thailand were expected to remain relatively stable. The economy in Thailand was still expected to grow in 2005 despite its tourist industry being damaged.

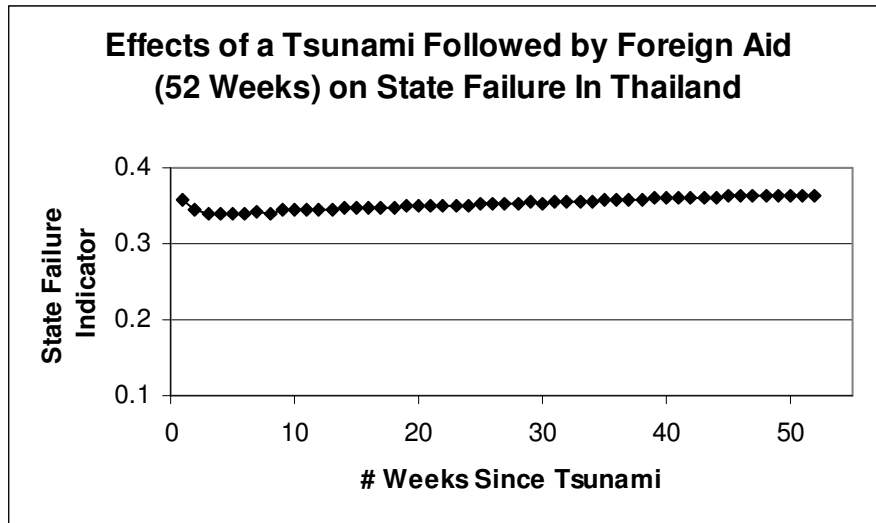
Figure 8 and Figure 9 show the output from the model that was used to compare the model predictions to the main effects of the tsunami.



**Figure 8. The model predicts that levels of corruption in Thailand following a tsunami and subsequent foreign aid causes a marginal increase in levels of corruption.**

<sup>5</sup> Transparency International assigned Thailand a score of 3.6 in 2004.

<sup>6</sup> Thailand has received an estimated . These figures from the Development Assistance Database (DAD) compiled by the United National Human Development Programme.



**Figure 9.** The model predicts that state failure is more likely to occur after a tsunami than before. The increase tapers off after about 6 months.

Regional experts on Thailand agreed that the level of corruption in the country would increase after the tsunami, though not immediately. Much less has been published about corruption from aid donations in Thailand than about Indonesia. This may be because Thailand was not affected to the same magnitude that Indonesia was. Consequently, much less aid flowed into Thailand than in Indonesia. The principle concern of corruption in Thailand is in the reconstruction phase where it is speculated that land once used as residential properties by the local population before the tsunami may be given to resorts. By wiping out residential neighborhoods where people had little proof of ownership of their land or property, it is harder for people to assert their right to the land they once occupied. In some areas, resorts are being constructed on land once occupied by poor residents.

The model's predictions show a slight increase in corruption after about four months which is consistent with expectations from the regional experts. However, evidence exists that the model may not capture all of the corruption or the type of corruption likely to be seen in Thailand. Evidence is growing that the businesses and government may be seeking to profit from the tsunami by allowing resorts to build on desirable tracts once inhabited by the local population<sup>7</sup>. The model dynamics for corruption assume that aid is grafted, that each instance of aid, be it through rehabilitation or reconstruction projects, results in a portion of that money being used for private gain. Consistent with the model, regional experts agreed that corruption would increase after a period of time, however, they assumed most of the corruption would result from skimming from aid received. Corruption derived from an opportunity that a disaster creates, such as in Thailand when an area of residential land is wiped out and can be used for resorts, is not captured by the model dynamics.

Similar to Indonesia, the model predicts a tsunami will have limited impact on the overall indicator for state failure in Thailand. A small dip in the value of the indicator occurs shortly after the tsunami and is due to an overall decrease in hostile activities across the country.

<sup>7</sup> See <http://www.csmonitor.com/2005/0408/p07s02-woap.html?s=hns> and <http://edition.cnn.com/2005/WORLD/asiapcf/12/04/tsunami.paradiselost.ap/>

Attempting replication can be valuable regardless of how well the model replicates the real-world. Successfully matching the real world in additional contexts is valuable because it helps convince potential users that the model can be generalized to different contexts. When considering using the model in another context that is deemed to be similar to contexts in which the model performed well, more confidence may be placed into using the model in this area.

Failing to replicate can be just as valuable as it provides evidence of where the model is not valid; it helps define the operational boundaries and can help developers identify the gaps in the conceptual model that lead to the replication attempt to fail. For example, with the RTE, the exercise of trying to replicate results in Thailand led the developers to discover a separate form of corruption not modeled in the RTE. While the model matched the expectations of the subject matter experts, both the model and the experts used similar assumptions.

### 6.2.3 Hypothetical

The fact that simulation models are addressing hypothetical questions means that there are constraints on their validation. Hypothetical questions involve thought experiments into conditions in the real-world that have not taken place, and thus real data does not exist. An inability to use real data to compare the model to means we must turn to others ways to evaluate the model's forecasts. An alternative is to evaluate the model's mechanisms and assumptions, which tends to be a subjective exercise of comparing the literature in which the model is based to the model's purpose, how appropriately it formalizes the theory, and whether the theories used are the correct ones. Experts in the domain can also be used to evaluate the model predictions. An additional method is to compare across models of similar purpose (Axtell 1995), checking the extent to which similar model parameterizations between two or more models result in similar predictions.

For the hypothetical experiment, terrorist activities were interjected in Thailand once weekly for 6 months. A basic insurgency style of attack was used with multiple small attacks, by the key terrorist group, primarily in the non-Islamic provinces and directed against primarily civilian targets. Regional experts expected that sustained terrorist activities would increase the likelihood of state-failure but only marginally, not enough to make the state weak or to fail. Figure 10 compares the overtime state-failure indicators of the sustained terrorist activity level scenario to the normal scenario which assumes no sustained terrorist activities. To check against the regional experts' expectations the magnitude of the state-failure indicator was compared between the two cases overtime. For the first six months while terrorist activities are on-going, the state-failure indicator in the terrorism case is greater than in the base case.



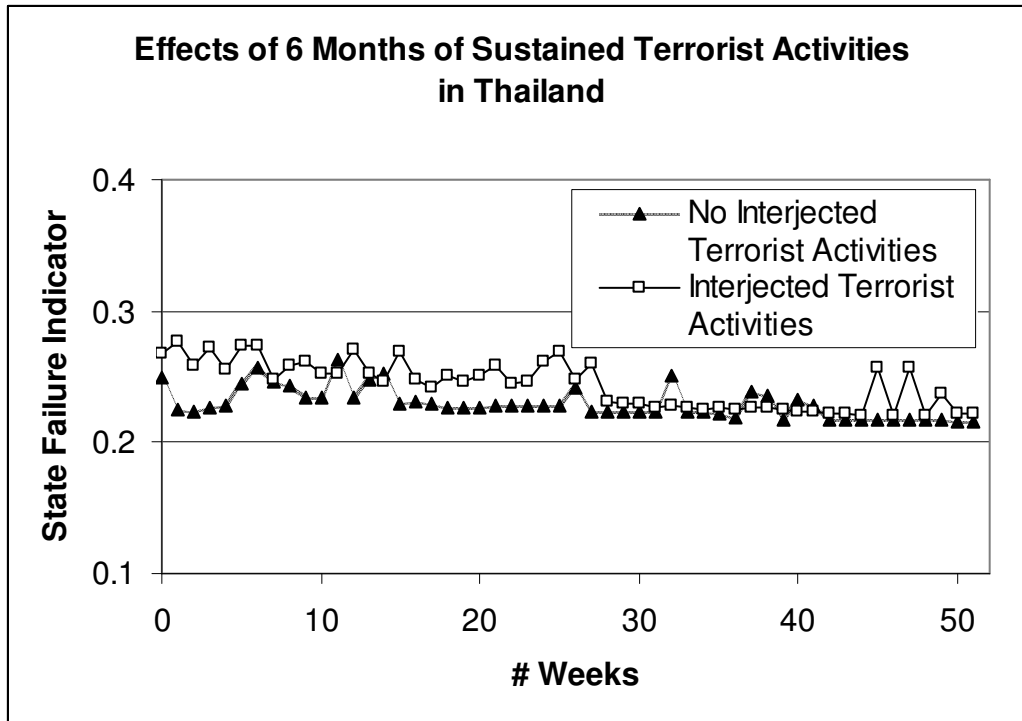


Figure 10. The effects of sustained terrorist activity in Thailand.

Evaluating how well RTE performed hypothetical analyses raised several issues. The subject matter experts who evaluated the RTE were experts on particular issues in Indonesia and Thailand. Example issues included women's right, terrorism, religious movements, and separatist movements. Subject matter experts used their domain knowledge about a few actors relevant to their domain expertise to extrapolate up to system-level behavior. Because the experts were knowledgeable about different issues, and consequently different actors, experts occasionally disagreed about the system-level effects. Evaluation of the RTE's predictions is complicated when one considers that the RTE accounts for all actors that are thought to be relevant whereas the subject matter experts tended to place emphasis on only a subset. Any difference between model output and expected system-level behavior could be because the subject matter experts have localized knowledge and have difficulties extrapolating.

Process-level comparisons to other models in the group were difficult to do because they had modeled different actors and different modeled each actor differently than is done in the RTE. Though system-level comparisons were still possible they were only able to be done at a very high-level

Comparisons to other models were limited to system-level predictions because each model used different actors and modeled the actors differently. Process-level comparison between models could only be done qualitatively because model outputs were different.

Despite the issues, the RTE can still be useful to study state stability dynamics. A multi-agent model should be thought of as formally representing a mental model. As such, the predictions that the model produces can be thought of as predictions that are a consequence of the mental model. When a multi-agent model is used to systematically explore a range of scenarios, its predictions can be compared to what is expected. When a mismatch occurs it can indicate that the mental model is flawed, but this will also require that the underlying process leading to the unexpected prediction be understood. Are assumptions violated? If agents are not

behaving as expected, what dynamics might be the cause? Are there factors not present in the model that need to be? This is not a complete list of questions that model developers need to ask when model results do not match the expected results, but they are indicative of the types of questions that need to be answered to better understand one's own mental model. Systematic hypothetical experimentation provides a structured framework to explore one's own mental model of a system.

#### **6.2.4 Forecasting**

Hypothetical experiments have the advantage of designing a range of possible scenarios to quickly test a model against the known theory and expectations of subject matter experts. The drawback is that it does not use data from the real-world to perform the evaluation, making it more difficult to ascertain how well the model relates to the real world. Parameterizing a model to represent the current day and forecasting forward provides a more stringent evaluation of a model.

For the RTE, in August 2005, two-year forecasts on the degree of state-failure were made for Indonesia and Thailand. By definition, forecasting is about making predictions about the future. However, it requires a sustained funding commitment beyond the time it takes to build and analyze the model. The issue is a practical one. Most funding is directed toward model development and validation using historical data, leaving little time and resources to collect data after the model has been sufficiently developed and evaluated using other validation types.

Some may wonder if forecasting can be simulated by splitting a historical data set into two and using one set to calibrate and the other to test, to simulate forecasting. Even without seeing the set used to test with, a developer will be likely influenced by what they know about what's already happened this will create bias when developing the model. In addition, declaring complete ignorance of the data is not likely to persuade anyone who is interested in learning about the validation method, no matter how truthful the declaration.

### **6.3 Data Validation**

We spend much less time discussing the extent to which the data is accurate, because this is a problem that most models face. One of the unique problems with models of socio-political systems is that input data into the model are often qualitative. This requires an interpretation of the qualitative data into valid ranges that the model works with. Most numerical values in the RTE range from 0 to 1 or -1 to 1. Thus, before collection begins descriptions of the extreme values must be given in a way that it is as unambiguous as possible how to interpret a qualitative description of the variable being collected. Consistency in data collection is important, so if multiple people are collecting data, inter-reliability checks should be made to measure and ensure consistency. Interpretation of a single data collector may even change over time as their experience in the area grows.

This corroborative data collection process is time-consuming and expensive. This is not grounds to dismiss the effort, but to recognize a real constraint on projects that are funded, that funding agencies may not want to pay for the additional resources to fund efforts to ensure data validity and instead focus on model development. This was the case with the project that funded the RTE and so formal inter-reliability checks have not been performed. We did provide stated

descriptions of all variables accompanied with interpretations of the extreme values. Appendix B gives a description of the variables used in the model.

## **7 Validating an Integrated Model of Nation-State Failure: Overview of the Issues**

### **7.1 Defining the Operating Domain**

Defining the operating domain of multi-agent models is difficult for at least two reasons: many assumptions are implicit and the parameter space is very large. Most multi-agent models are programmed using a special modeling language or a more general language such as C++. Regardless of how it is implemented, the assumptions of the model are embedded in the code. In contrast to a regression model where the assumptions of a regression model are known when the structure is defined, the construction of a multi-agent model requires that assumptions be clearly stated outside of the model or that those interested in assumptions read code. Stating the assumptions outside of the code requires basically writing another document that is as long or longer than the code itself. In writing up such assumptions, as in writing up a verbal theory, it is easy to inadvertently omit assumptions. Checking assumptions is difficult because it requires that the code or a suitable representation of the code be made available. However, making the code available may invalidate IP claims.

Large parameter spaces, characteristic of multi-agent models, permit a potentially large response surface. The challenge then becomes determining over which ranges and sets of parameters the model is capable of producing valid results. In the case of the RTE, sensitivity analysis and replication helped to define how far out the current structure of the model can be pushed and still get answers that approximate the real-world. For the RTE sensitivity analysis showed that the level of corruption before foreign aid starts to flow affects how quickly corruption increases when aid is received. However, it also showed that lower initial levels of corruption resulted in steeper increases in corruption over time, which is not what we might expect especially when corruption levels are very low. In these cases corruption should not increase or increase only marginally since we would expect uncorrupt governments not to skim from received aid. Sensitivity analysis of the RTE helped to set a boundary by demonstrating that the model may not operate correctly when the government is not corrupt.

Doing this analysis while informative, only covered a small portion of the response surface. This is a typical problem for multi-agent models. In general one needs to use data-farming techniques to fully evaluate the response surface. Even then, given the size of typical parameter spaces there may not be sufficient computer storage space for the results from a comprehensive analysis. In addition, data-farming environments, such as that at the Maui high performance computer center are not easily nor routinely available to most researchers and require that the model be written with certain web enabling features.

Replication provided an additional method for defining the operating domain. Dynamics for corruption were developed using what was known about Indonesia. Though the RTE matched regional experts' beliefs about corruption in Indonesia and what is known through published reports, it may not appropriately replicate what is occurring in Thailand. Corruption in Indonesia is assumed to stem from aid flowing into the country whereas corruption in Thailand might also occur over land disputes between resort owners and the local population. The latter cause of corruption is not represented in the RTE, clearly marking a boundary for where the

model can accurately represent corruption dynamics. In general, replication is a powerful and effective technique from a time and space constraint perspective. However, it requires both multiple subject matter experts and multiple modeling teams. Given tight resources, having such multiplicity is often viewed as redundant and a waste of resources by funding agents; rather, than as a necessary component of validation.

## **7.2 Applicable Techniques of Validation**

The availability and types of data constrain what kinds of validation techniques can be applied. For the RTE, reports of levels of corruption are hard to come by, though it is suspected that it occurs in Indonesia and Thailand. Often, aid agencies responsible for collecting and disbursing aid are reluctant to report misuse of funds because acknowledgement of its occurrence may reduce the amount of aid donated. Consequently, the model's output can only be compared to reports, news sources, and subject matter experts. A key problem from a validation angle is that these sources may, and often do, disagree. Thus the most that can be done is to show the conditions under which the model matches the different points of view or, as is done in expert systems, that it matches some weighted average of these alternative points of view.

## **7.3 Theory Integration and Validation by Parts**

Most scientific theories have a *ceteris paribus* assumption, that the theory is valid while other factors are constant. It is a necessary assumption to have in order to rule out possible other relevant factors and determine the relative effect of a single factor. This assumption is at odds with multi-agent systems where part of the expressiveness of the model is derived from the interaction of multiple factors. In multi-agent systems it is possible to integrate multiple theories to drive the dynamics. While each of these theories can go through separate validation procedures where the *ceteris paribus* assumption holds, integrating them other theories presents the possibility that interactions between the theories can produce unforeseen results.

Once the individual parts are combined, the open question is whether the integrated model is also valid. The RTE draws from separate theories of state failure processes where the theories do not explicitly consider other processes, and thus the extent to which the complete model is valid from its parts is an issue.

# **8 Conclusions**

We find that various features of multi-agent models call into question traditional validation approaches. For example, large scale multi-agent systems have high dimensionality, possibly higher than that of the data available for validation. Designs of the models are not axiomatic in the sense that the rules and behavior need to follow any formal mathematical constraints, making assumptions difficult to express. These challenges have implications for how these models ought to be used. For example, are they most useful for proving concepts, generating new hypotheses to test using alternate methods, or formulating policy recommendations? These challenges also have implications for how these models can be validated. Nevertheless, the appropriate use of these models will depend on a number of subjective factors, including, but not limited to, the kind of validity that can be achieved.

Statistical techniques that are available to more traditional engineering models are not available to many multi-agent models. Instead, output from the model is compared to expected qualitative features of the output. When performing sensitivity analysis, calibration, and

replication, and forecasting the developers can inspect the model's output and compare it to what is known in the literature. When hypothetical analyses are performed or when an extant literature on a phenomenon does not exist, subject matter experts must be used to validate the model output. A reliance on subject matter experts means the model is being compared to the mental models of the experts. Mental models can be flawed for a number of reasons and the mental models between experts may be different resulting in conflicting evaluations of a model.

Though challenges exist in validating multi-agent models, confidence can still be developed in these types of models, but should follow some guidelines. First, the operating domain needs to be defined. Conceptual validation through sensitivity analysis can help reveal in what parameter spaces the model does not operate faithfully to the real-world. Replication of additional contexts beyond what the model was originally calibrated for, can reveal assumptions of the model and uncover processes the model cannot explain. Second, assumptions need to be explicitly stated. This can be done by stating variables used, the equations used, and describing the main simulation loops of the model.

Even supposing very limited validation of a multi-agent model, they can still serve a role in a policy context. The development of the model forces assumptions to be revealed, and how the system is chosen to be modeled forces consideration of what the important parts of the system are and at what level they should be modeled. Thus, constructing a multi-agent model provides a method to formalize mental models of a real-world system. Once constructed, the multi-agent model becomes a formal representation of a mental model. The model can then serve to corroborate or question existing beliefs or reveal possible important scenarios that were not previously considered. Given errors people are prone to have in their mental models, formalizing them and then pushing to see the consequences of their mental model allows the discovery of inconsistent or wrong beliefs.

Throughout this paper, we used the RTE as an example of a dynamic-network multi-agent model. The purpose of the model is to explore conditions that influence indicators of nation-state failure. Due to the complexity of the model, validation techniques were limited to comparisons to extant theory, subject matter experts on Indonesia and Thailand, published reports, and news sources. Nonetheless, the validation process was able to help define the operating domain. The RTE's role in a policy is best-suited as a tool to help an analyst or policy maker explore the space of relevant policy options that might affect indicators of state failure.

## References

- Action Contre La Faim* (2005). *Preliminary Analysis on the Food Aid response to the tsunami crisis*, Action Contra La Faim.  
[http://www.odi.org.uk/hpg/papers/ACF\\_analysis\\_aceh.pdf](http://www.odi.org.uk/hpg/papers/ACF_analysis_aceh.pdf).
- Graham T. Allison and Philip Zelikow (1999). *Essence of Decision: Explaining the Cuban Missile Crisis*, Longman.
- Robert Axelrod (1995). *The Convergence and Stability of Cultures: Local Convergence and Global Polarization*. Santa Fe, NM, Santa Fe Institute: 95-03-028.
- Robert Axtell, Axelrod, Robert, Epstein, Joshua M., Cohen, Michael D. (1995). "Aligning Simulation Models: A Case Study and Results." *Computational & Mathematical Organization Theory* 1(2): 123 - 142.

- J. Banks and D. Gerstein and S.P. Searles (1987). Modeling Processes, Validation, and Verification of Complex Simulations: A Survey. SCS Simulators Conference, Orlando, FL, The Society for Computer Simulation
- Emma Batha (2005). *Q & A: Corruption and aid*, ReliefWeb. 2005.  
<http://www.reliefweb.int/rw/RWB.NSF/db900SID/SODA-6J259E?OpenDocument>.
- Markus Bergkvist, Davidson, Paul, Persson, Jan A., Ramstedt, Linda (2004). A Hybrid Micro-Simulator for Determining the Effects of Governmental Control Policies on Transport Chains. Joint Workshop on Multi-Agent and Multi-Agent Based Simulation, New York, NY, Springer
- James H. Bigelow, Davis, Paul K. (2003). *Implications for Model Validation of Multiresolution, Multiperspective Modeling (MRMPM) and Exploratory Analysis*. Santa Monica, CA, RAND.
- Peter M. Blau (1974). "Presidential Address: Parameters of Social Structure." American Sociological Review 39: 615-635.
- Peter M. Blau (1977). Inequality and Heterogeneity: A Primitive Theory of Social Structure. New York, The Free Press of Macmillan Co.
- Richard M. Burton (2003). "Computational Laboratories for Organization Science: Questions, Validity and Docking." Computational & Mathematical Organization Theory 9(2): 91-108.
- Richard M. Burton and Børge Obel (1995). "The validity of computational models in organization science: From model realism to purpose of the model." Computational & Mathematical Organization Theory 1(1): 57 - 71.
- Kathleen M. Carley (1991). "A Theory of Group Stability." American Sociological Review 56: 331-354.
- Kathleen M. Carley and Douglas B. Fridsma and Elizabeth Casman and Alex Yahja and Neal Altman and Li-Chiou Chen and Boris Kaminsky and Démian Nave (2006). "BioWar: Scalable Agent-Based Model of Bioattacks." IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 36(2).
- Jeffrey S.; Gumerman Dean, George J.; Epstein, Joshua M.; Axtell, Robert; Swedlund, Alan C. (2000). Understanding Anasazi Culture Change Through Agent-Based Modeling. Dynamics in human and primate societies : agent-based modeling of social and spatial processes. Timothy; Gumerman Kohler, George J. New York, Oxford University Press.
- Jack David Eller (1999). From Culture to Ethnicity Conflict. Ann Arbor, The University of Michigan Press.
- Joshua M. Epstein and Robert Axtell (1996). Growing Artificial Societies: Social Science From the Bottom Up. Cambridge, MA, MIT Press.
- Joshua Epstein and John D. Steinbruner and Miles T. Parker (2001). Modeling Civil Violence: An Agent-Based Computational Approach. Washington, D.C., Center of Social and Economic Dynamics, Brookings Institute.
- Leon Festinger (1954). "A Theory of Social Comparison Processes." Human Relations 7: 117-140.
- Jay W. Forrester (1961). Industrial Dynamics. Cambridge, MA, MIT Press.
- Jay W. Forrester (1969). Urban Dynamics. Cambridge, MA, MIT Press.
- Jay W. Forrester (1971). World dynamics. Cambridge, MA, MIT Press.
- Noah Friedkin (1998). A Structural Theory of Social Influence. New York, Cambridge University Press.

- Noah Friedkin (2003). *Toward a Science of Strategic Modification of Interpersonal Influence Systems. Dynamic Social Network Modeling and Analysis*. Kathlen Carley Ronald Breiger, and Philippa Pattison. Washington D.C., National Academy of Sciences/National Research Council Committee on Human Factors: 89-100.
- Noah E. Friedkin (1998). *A Structural Theory of Social Influence*. New York, Cambridge University Press.
- Jack A. Goldstone and Ted Robert Gurr and Barbara Harff and Marc A. Levy and Monty G. Marshall and Robert H. Bates and David L. Epstein and Colin H. Kahl and Pamela T. Surko and John C. Ulfelder and Alan N. Unger (2000). *State Failure Task Force Report: Phase III Findings*. McLean, VA, Science Applications International Corporation.  
<http://www.cidcm.umd.edu/inscr/stfail/>.
- Donald L. Horowitz (1985). *Ethnic Groups in Conflict*. Berkeley, University of California Press.
- Kirsten Johnson (2004). *Political Capacity and the Severity of Internal Conflict: Applying Relative Political Extraction to Ethnic Conflicts*. Journeys in World Politics Workshop, University of Iowa
- W. O. Kermack and A. G. McKendrick (1927). "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society* **115**: 700-721.
- Jack P.C. Kleijnen (1999). *Validation of models: statistical techniques and data availability*. *Proceedings of the 1999 Winter Simulation Conference*, Phoenix, AZ
- Averill M. Law and David W. Kelton (1999). *Simulation Modeling and Analysis*. New York, NY, McGraw-Hill.
- Lesley McCulloh (2005). *Aceh: Then and Now*, Minority Rights Group International.
- Jamison Jo Medby and Russel W. Glenn (2002). *Street Smart: Intelligence Preparation of the Battlefield for Urban Operations*, RAND Corporation.
- E.J. Mishan (1973). *Economics for Social Decisions: Elements of Cost-Benefit Analysis*. New York, New York, Praeger.
- E.J. Mishan (1973). *Elements of Cost-Benefit Analysis*. New York, New York, Praeger.
- Scott Moss and Bruce Edmonds (2005). "Sociology and Simulation: Statistical and Qualitative Cross-Validation." *American Journal of Sociology* **110**(4): 1095 - 1131.
- Jeffrey Pfeffer, Salancik, Gerald R. (1978). *The external control of organizations: A resource dependence perspective*. New York, Harper and Row.
- Robert G. Sargent (1992). *Validation and Verification of Simulation Models*. 1992 Winter Simulation Conference, Piscataway, New Jersey, Institute of Electrical and Electronics Engineers
- Leonard Simanjuntak (2005). Interview with Leonard Simanjuntak. Berlin, Germany, Transparency International. **2005**.
- Herbert Simon (1955). "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* **69**: 99-118.
- Herbert A. Simon (1982). *Models of Bounded Rationality*. Cambridge, MA, MIT Press.
- Neill Stansbury (2005). *Exposing the foundations of corruption in construction*. Berlin, Transparency International: 36-40.
- Jan Thomsen and Raymond E. Levitt and John C. Kunz and Clifford I. Nass and Douglas B. Fridsma (1999). "A Trajectory for Validating Computational Emulation Models of Organizations." *Computational & Mathematical Organization Theory* **5**(4): 385 - 401.

*Alex Yahja and Kathleen M. Carley (2006). WIZER: Automated Model Improvement in Multi-Agent Social-Network Systems. Coordination of Large-Scale Multiagent Systems. Paul; Vincent Scerri, Regis; Mailler, Roger, Springer: 255 - 270.*

## **Appendix A: Micro-Theories in the RTE**

The RTE is a discrete time-based model. Figure A.1 gives a high-level view of the main simulation loop. At the start of time period, each agent begins by deciding whether to take an action, shown by the top-most node in the diagram. If they do, the simulation follows the arrow to right marked “Yes,” and otherwise it follows the arrow marked “No.” The nodes in the diagram represent the main behavior of the program, and each of these and their groundings will be discussed in this section.

Agents vary in level (entity, province, and state), type (NGO, government, military, corporate, etc.), nature (red, blue and green), tension, tendency to take risks, historical activity level, goals and level of resources. Goals are defined in terms of preferences for social, symbolic or economic effects. These agents, each time period, decide whether or not the situation warrants them taking action, then if it does they choose both an action and a target, and then take that action. In RTE all agents act effectively in parallel. So these actions can be at odds with each other. For these agents a time tick represents a week of real time. Once an agent has taken an action, that action consumes various resources on the part of the agent and the target, impact the tension of the agent and impacted agents, and alter the influence of the agent on others and their influence on the agent. Agents can engage in multiple actions at once.

Agents are connected into a set of networks. These include the influence network that determines which agents affect others action choices and the hostility/non-hostility network that determines the type and direction of action one agent takes on another. Influence is a function of proximity, socio-demographic similarity, resource levels, and historical influence. Hostility/non-hostility is historically based. Both of these networks evolve over time. As agents take hostile acts toward each other the level of hostility increases whereas taking a non-hostile act decreases the level of hostility. As agents do not follow the “advice” of those who have influence over them, the influence of those others decreases. So disagreement lowers influence and agreement tends to increase it.

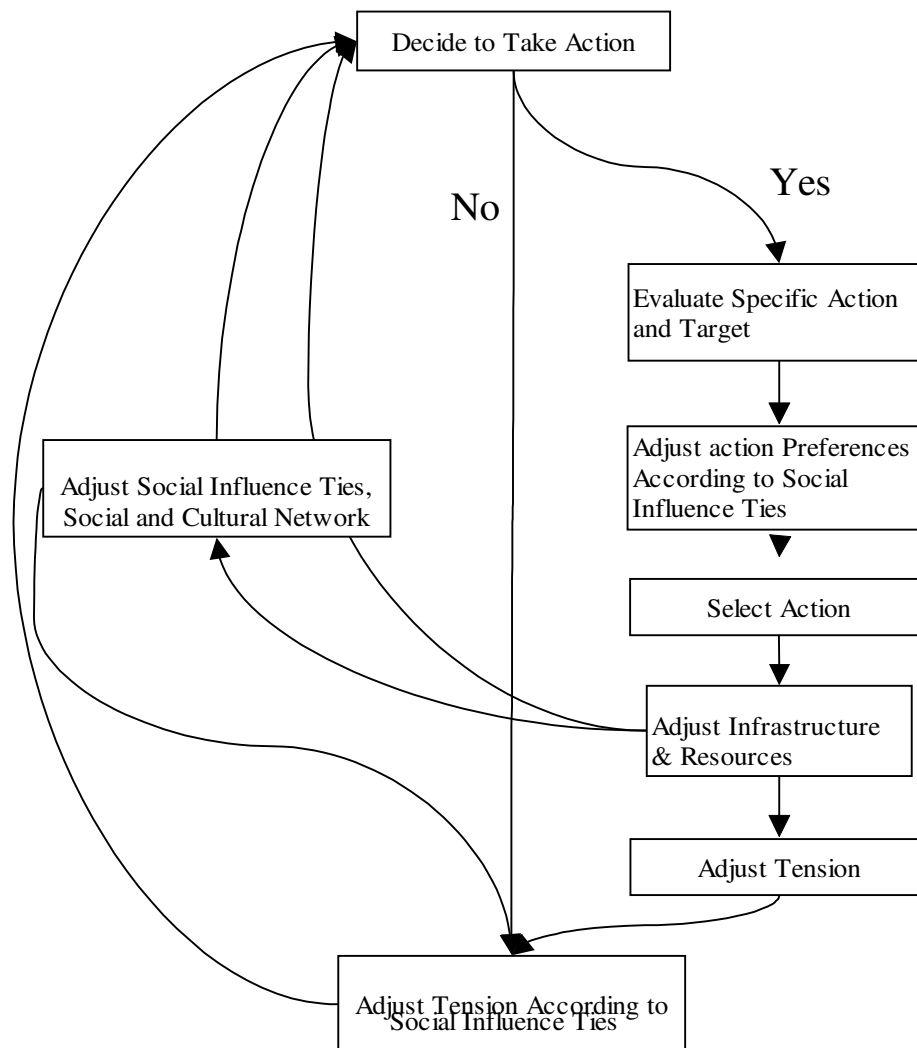
The actions taken by the agents vary in type, direction, resources consumed, damage generated, social, symbolic and economic impact, the level of physical, planning and resource effort needed to take an action, and the number of time-periods for which they last. The types of action are military, political/diplomatic, social, economic, information, infrastructure, and criminal. Action directions are hostile, neutral or friendly. Strength is measured on a three point scale – low, medium, high. So a low hostile political action does less damage to the targets social resources on average than a high hostile military action. Actions can be directed to another agent or toward a physical target e.g., radio-station. Hostile actions tend to increase tension and non-hostile actions lower tension. Actions can also consume or generate resources for the agent or a targeted agent. The impact of each action on the agents and the state and province indicators is implemented using a series of weight and adjustment rules.

Agents decide each time period to take action. This is modeled using a social influence model in which the desire to take an action is a function of both the level of tension and the influence by others encouraging or discouraging the taking of action. If an action is to be taken the agent then selects an action and a target using a cost-benefit calculation modified to account



for both resources and rationality bounds; i.e., agents cannot take actions or attack targets that require substantially more resources than they have and not all options are evaluated. This cost-benefit calculation takes into account the combined potential social, symbolic and economic impact and the planning, resource, and physical effort needed. This results in a preference for an action and target by that agent. This preference is then modified using a social influence module to account for the social influence of other agents on what action this agent should take and what the target should be. Agents, when “giving advice” to the acting agent use their opinion about the impact and effort required by the actor and they may be wrong because they have a flawed understanding of the capabilities of the actor.

For example, if agent a takes a hostile action on agent b that will decrease a’s tension and increase b’s. In addition, if agent c is socially influenced by or proximal to b, then c’s tension will also increase using the social influence model.



**Figure A.11.** Conceptual picture of the dynamics of a single agent during each time period.

## A.1 Deciding to Take Action

Agents are characterized by a concept of tension, which is a way to describe how unsettled, anxious, or aggravated a group of people who the agent represents feel. Factors that indicate levels of tension include whether they are in conflict with one or more groups (Horowitz 1985; Eller 1999), how much disparity exists in income and access to resources (Blau 1974), and what the health conditions are. These are used to help set initial tension levels in agents. The tension an agent feels, the more they sense a problem exists, and the more likely they are to make a problemistic search (Simon 1955) to address the problem

## A.2 Individual Action Selection

The decision-making framework is based on a deterministic benefit-cost analysis (Mishan 1973). Agents weigh the perceived impacts (benefits) of an action versus how difficult the action is to perform (costs). The benefits of an action are perceived along three different dimensions, 1) economic, 2) symbolic, and 3) social. While the agents are assumed to perceive the benefits and the costs of the actions they are going to take, the selection of the action and the target (what they take an action on) is stochastic. When the weighted difficulty is subtracted from the weighted impact, the resulting score is a weighted attractiveness of an action by a particular agent. Let  $a_{xkj}(0)$  represent the initial perceived attractiveness of action-task  $k$  to agent  $i$  as perceived by himself.

$$a_{iik}(0) = \text{Impact}_{iik} - \text{Difficulty}_{iik}$$

**Equation 2.** Initial perceived attractiveness of an action-task  $k$  for some agent  $i$ .

Difficulty is further described along three different dimensions. Traditionally, task difficulty has been described in terms of planning time ( $T$ ), the amount of resources needed to complete the task ( $R$ ), and the physical difficulty in performing the task ( $P$ ). These notions are used to describe the difficulty of the task. The difficulty component is weighted by  $S_d$ , which describes the relative skill level of the agent. Higher values of  $S_d$  indicate more skill.

$$\text{Difficulty}_{iik} = (1 - S_{d,i}) \frac{(T_{iik} + R_{iik} + P_{iik})}{3}$$

**Equation 3.** The difficulty component separated into its constituent variables.

The expected impact that an action has is described as the weighted sum of an action's ability to cause economic impact (damage or provision of resources) ( $I_e$ ), social impact (casualties or improved social welfare) ( $I_c$ ), and symbolic impact ( $I_s$ ). Symbolic impact includes psychological damage and psychological support. The weights describe an agent's preference for causing economic damage ( $E$ ), casualties ( $C$ ), or symbolic damage ( $S$ ). Although,  $I_e$ ,  $I_c$ ,  $I_s$  are -1 to 1, total impact is the absolute value of the sum.

$$\text{Impact}_{iik} = \frac{E(I_{E,iik}) + C(I_{C,iik}) + S(I_{S,iik})}{3}$$

**Equation 4.** The impact component and its constituent variables.

Explicitly, the initial attractiveness of a particular action-task is calculated using the following equation.

$$a_{iik}(0) = \frac{E(I_{e,iik}) + C(I_{c,iik}) + S(I_{s,iik})}{3} - (1 - S_d) \frac{(T_{iik} + R_{iik} + P_{iik})}{3}$$

**Equation 5.** Complete equation used to calculate the attractiveness of action-task pairs.

Initial attractiveness ranges from -1.0 to 1.0. If the attractiveness score is negative, then the agent is assumed not to consider performing the action at all.

The initial attractiveness of a specific action-task for an agent, ego, perceived by an agent, self, does not consider the influence of other agents. This will be addressed in the next section in determining the final attractiveness. We assume that the hostile agent randomly chooses the action-task in proportion to the attractiveness of the action-tasks. This assumption will be relaxed in stage 2, in a future iteration of the model.

### A.3 Social Influence on Action Selection – Creation of the Final Attractiveness

Agents in the system also have ties to other agents in the system. Who an agent is tied to and the nature of those relationships defines how much influence others have on their decisions. Social influence in the model is formalized by using Noah Friedkin's model of social influence (Friedkin 1998; Friedkin 2003). It includes the set of each agent's initial attack preference, the influence ties between agents, and the susceptibility each agent has to being influenced.

When considering an attack, an agent does not only consider their own preferences, but also the preferences of those who have influence over them. Consider a system of  $N$  agents. Using the equation for attractiveness given in section 1.1, let  $a_{iik}$  be the initial attractiveness of a specific attack choice  $k$  for agent  $i$ . To add the influence of others in the attack choice, the following formulation is used.

$$a_{iik}(1) = s_i(w_{i1} a_{i1k}(0) + w_{i2} a_{i2k}(0) + \dots + w_{in} a_{ink}(0)) + (1 - s_i) a_{iik}(0)$$

**Equation 6.** The attractiveness of a specific action type for an agent using both the individual preference and the social influence of others.

The influence of others is moderated by weights which describe the degree of influence each agent has on another agent. The weights in the model range from 0 to 1.0. The susceptibility of the agent  $i$  to being influenced is represented by the term  $s_i$ . It also ranges in value from 0 to 1.0. Action-tasks are selected to be used probabilistically based on their attractiveness scores. Each action-task receives an attractiveness score. One of these are randomly chosen, weighted by their attractiveness. The probability that the agent actually performs the task is also based on the selected action-task's attractiveness.

### A.4 Dynamics of Social Influence

After each time period, the nature of relationships change according to what actions were taken in the time period. Influence relationships change according to two different ways. The first way is according to the difference in resources. Agents who have more resources are assumed to be more influential. Resources are used during each time period, requiring that the influence relationships be updated. Let  $w_{ij}$  be the influence that agent  $j$  has over agent  $i$  and let  $R_i$  and  $R_j$  represent the resources that an agent  $i$  and  $j$  have respectively. (Note: if  $w_{ij} \leq 0$ , then agent  $j$  has no influence over  $i$ .)

$$w_{i,j}^{(t+1)} = (R_j - R_i)$$

**Equation 7.** Resource-based influence relations among agents.

Influence also changes according to the perception of what other agents should be doing. An agent  $y$  has a perception of how attractive an action-task is to other agents whom they have relationships with. For each other agent  $x$  and action-task  $z$ , agent  $y$  has a perceived  $a_{xyz}$ . An agent  $y$  believes that another agent  $x$  should pursue an action-task that the former believes to be the most attractive to  $x$ . The preference of agent  $y$  for agent  $x$  to perform an action,  $a_{xy}$ , is defined as

$$PA_{xy} = \max_z PAA_{xyz}$$

**Equation 8.** How beliefs of what other agents should be doing are developed.

When the action-task  $z$  that maximizes  $PAA_{xyz}$  is not the same as the task  $j$  that maximizes  $PAA_{xxz}$  and agent  $y$  has influence over  $x$ , then the influence that  $y$  has over  $x$  is adjusted in the model to reflect the evidence that  $y$  was not able to influence  $x$ .

$$w_{i,j}^{(t+1)} = w_{i,j}^t \cdot (1 - s_i)$$

**Equation 9.** Influence adjustment for the next time period when an agent was unable to influence another agent

## A.5 Weight and Adjustment

Variables in the model are modified via a weight and adjustment formula designed to keep the values between 0 and 1 or between -1 and 1. How quickly variable values increase or decrease is modified by an adjustment factor and a volatility factor. The adjustment factor is associated with a particular event or condition in the environment. Different adjustment values for different events and conditions specify the relative influence an event or condition has on an agent. Each agent has an adjustment factor for each event and condition that can occur in the system. Different agents may use different adjustment factors.

Whereas the adjustment factor is associated with an event or condition as well as an agent, the volatility factor is associated with only the agent. The volatility factor specifies the degree to which an agent is affected by events and conditions in general.

The structure of the weight and adjustment formula for variables that vary between 0 and 1 for a specific agent or sector is given as

$$T_{i,t+1} = T_{i,t} + (1 - T_{i,t}) \cdot A_{i,k} \cdot V_i$$

Equation 10. Weight and adjustment formula used to update tension levels for sectors and agents.

where  $A_{i,k}$  is agent  $i$ 's adjustment factor associated with an event or condition  $k$ ,  $V_i$  is the agent's volatility factor, and  $T$  is the variable of interest for the agent. Every pair  $\langle i, k \rangle$  is associated with three adjustment factors depending on the level of "success" (randomly chosen) of the event. Low success actions adjust tension more slowly than highly successful actions of the same type.

The structure of the weight and adjustment formula for variables that vary between -1 and 1 is given as

$$T_{i,t+1} = T_{i,t} + (1 - |T_{i,t}|) \cdot T_{i,t} \cdot A_{i,k} \cdot V_i$$

## A.6 Social Influence on Tension Adjustment

Associations with others who are tenser than we are tend to make us tenser. Similarly, those who are less tense than us tend to calm us. Using Friedkin's theory of social influence (Friedkin 1998; Friedkin 2003) on beliefs is a useful way to formalize this notion. It is also important to capture, since tension is the main variable driving the whether any action is taken during a time period.

In a system of  $N$  agents, let  $T_i'$  be the updated tension of agent  $i$  and let  $T_i$  be its tension level before updating. The social influence that  $j$  has on agent  $i$  is expressed as  $w_{ij}$ . An agent  $i$ 's own susceptibility to being influenced is given by  $s_i$ .

$$T_i' = s_i(w_{i1} T_1 + w_{i2} T_2 + \dots + w_{in} T_n) + (1 - s_i)(T_i)$$

Equation 11. Social influence equation used to update tension in sectors and agents based on their social influence ties.

## Appendix B: Model Parameters

The purpose of this section is to describe the model parameters in the RTE.

### B.1 Entity Variables

Variable Name	Description
Default Economic Impact	The default perceived economic impact of an action taken against the entity
Default Economic Resources Required	The default perceived amount of financial/infrastructure resources required to perform an action against the entity
Default Perceived Physical Difficulty	The default perceived level of difficulty in mounting an action toward the entity
Default Perceived Planning Time	The default perceived amount of time it takes to prepare for an action against the entity
Default Social Impact	The default perceived impact on human capital of an action taken against the entity
Default Social Resources Required	The default perceived amount of human resources required to perform a hostile action against the entity
Default Symbolic Impact	The default perceived religious, political, or cultural impact of an action taken against the entity

Default Symbolic Resources Required	The default perceived amount of political, cultural or religious resources necessary to perform an action against the entity
Dominant Ethnicity	The predominate ethnicity of members of this group.
Dominant Language	The language predominately spoken by members of this group.
Dominant Religion	The religion predominately adhered to by members of this group.
Economic Resource Recovery Rate	The rate at which the state is able to increase or replenish economic resources
Initial Aggression Level	The tendency the entity has to act at the start of the simulation.
Initial Economic Resources	Financial and infrastructure resources
Initial Social Resources	Human capital, both in terms of size and quality/capabilities
Initial Symbolic Resources	Political, cultural, and religious resources
Initial Tension Level	The degree to which this group feels unsettled, worried, anxious, angry about the current state or what is to come
Preference for Causing Casualties (Social) Damage	The inclination that the entity has to affect the social well-being of people, either by causing or preventing casualties or deaths.
Preference for Causing Economic Damage	The inclination the entity has to affect economic activity
Preference for Causing Symbolic Damage	The inclination that the entity has to make symbolic statements through their actions
Relative Skill Level	How much experience does this entity have at performing their actions? How good are they, in general, at doing them?
Social Resource Recovery Rate	The rate at which the state can increase or replenish its human resources
Symbolic Resource Recovery Rate	The rate at which the state can increase or replenish its symbolic resources
Trust in State	How much trust does this group have in the state to provide them services (including human security) and use resources properly.

Type	Choose from: StateActor,USMilitary,ForeignMilitary,Green,Terrorist,InStateNonMilitary,InStateMilitary
Volatility	In general how much does this entity change its demeanor in the face of events of interest?

## B.2 Provinces

Variable Name	Description
Dominant Religion	Which religious group has the most adherents
Dominant Ethnic Group	Which ethnic group makes up the largest proportion of the population in the province
Dominant Language	The language spoken by the largest proportion of the population
Population	The total number of people living in the province
Population Density	The total number of people living in the province divided by the area (in km <sup>2</sup> )
Default Economic Impact	The default perceived economic impact of an action taken against the province
Default Symbolic Impact	The default perceived religious, political, or cultural impact of an action taken against the province
Default Social Impact	The default perceived impact on human capital of an action taken against the province
Default Perceived Planning Time	The default perceived amount of time it takes to prepare for an action against the province
Default Perceived Physical Difficulty	The default perceived level of difficulty in mounting an action toward the province
Default Economic Resources Required	The default perceived amount of financial/infrastructure resources required to perform an action against the province
Default Social Resources Required	The default perceived amount of human resources required to perform a hostile action against the province
Default Symbolic Resources Required	The default perceived amount of political, cultural or religious resources necessary to perform an action against the province
Initial Social Resources	Human capital, both in terms of size and quality/capabilities
Initial Economic Resources	Financial and infrastructure resources
Initial Symbolic Resources	Political, cultural, and religious resources
Social Resource Recovery Rate	The rate at which the state can increase or replenish its human resources
Economic Resource Recovery Rate	The rate at which the state is able to increase or replenish economic resources
Symbolic Resource Recovery Rate	The rate at which the state can increase or replenish its symbolic resources

Volatility	The degree to which conditions in the province tend to remain stable or are erratic
Initial Tension Level	The general level of unrest, fear, dissatisfaction among the population
Provision to State	The amount of resources flowing to the state in terms of taxes or natural resource extraction
Support from State	The amount of resources flowing into the province from the state (e.g., ROI on taxes)
Level of Criminal Activity	The degree to which conditions in the state tend to remain stable or are erratic
Activity Level of Foreign Military	How often foreign militaries operate inside of the country within a given time frame.
Level of Human / Women's Rights	How well are the rights of women and people in general respected.
Clean Water	How much access do people in the province have access to clean water
Food	How much access do people in the province have access to food
Medical/Health Care	How much access do people in the province have access to medical care
Train Infrastructure	The ability of the train infrastructure to move people across the province and to other provinces.
Road Infrastructure	The ability of the road infrastructure to move people across the province and to other provinces.
Air Transportation	The ability of the air transport infrastructure to move people across the province and/or to other provinces.
Degree of Radio Penetration	The degree to which radio can be used as a means of sending information to people. Considers both broadcast coverage and the number of people who use radios.
Degree of Television Penetration	The degree to which television can be used as a means of sending information to people. Considers the number of people who use televisions.
Degree of Internet Penetration	The degree to which the internet can be used as a means of sending information to people. Considers the number of people who have internet access
Elementary Enrollment	The degree to which the province provides education to young children.

### B.3 State

Variable Name	Meaning
Population Size	The size of the population of the country
Population Density	The population density of the country
Dominant Language	The language spoken by the largest proportion of the population
Dominant Ethnic Group	Which ethnic group makes up the largest proportion of the population
Dominant Religion	Which religious group has the most adherents
Socio-Economic Level	The level of economic wealth and social services delivered to the population



Media Penetration	The degree to which the population has access to mass media
Human/Women's Rights	The degree to which the state respects human rights and gender rights
Currently in War	Is the country's military currently engaged in a war or conflict
Foreign Military Activity	Presence and scope of actions of foreign military in this country
Frequency of Diplomatic or Peaceful Overtures	The frequency of stability-promoting actions and events
Frequency of Hostile/Lawless Actions	The frequency of destabilizing actions and events
Initial Tension Level	The general level of unrest, fear, dissatisfaction among the population
Volatility	The degree to which conditions in the state tend to remain stable or are erratic
Initial Economic Resources	Financial and infrastructure resources
Initial Social Resources	Human capital, both in terms of size and quality/capabilities
Initial Symbolic Resources	Political, cultural, and religious resources
Economic Resource Recovery Rate	The rate at which the state is able to increase or replenish economic resources
Social Resource Recovery Rate	The rate at which the state can increase or replenish its human resources
Symbolic Resource Recovery Rate	The rate at which the state can increase or replenish its symbolic resources
Default Perceived Physical Difficulty	The default perceived level of difficulty in mounting an action toward the state as a whole
Default Perceived Planning Time	The default perceived amount of time it takes to prepare for an action against the state
Default Economic Resources Required	The default perceived amount of financial/infrastructure resources required to perform an action against the state
Default Social Resources Required	The default perceived amount of human resources required to perform a hostile action against the state
Default Symbolic Resources Required	The default perceived amount of political, cultural or religious resources necessary to perform an action against the state
Default Economic Impact	The default perceived economic impact of an action taken against the state as a whole.
Default Social Impact	The default perceived impact on human capital of an action taken against the state.
Default Symbolic Impact	The default perceived religious, political, or cultural impact of an action taken against the state as a whole

## Appendix C: Parameter Space Constraints

Multi-agent models can have very large parameter spaces, which can make validating the model difficult or impossible because of the limited amount of data. The real world, however, places constraints on what are valid ranges of data. The constraints can collapse the real size of the parameter space.

For example, in the RTE both Indonesia and Thailand used 32 different agents. The agents were categorized into one of six different groups, part of the state military, part of state (non-military), sub-population group, terrorist group, and a foreign organization. Rather than parameterizing each agent separately, the agents were parameterized according to their classification into one these groups.

The space was further reduced by turning constraining all numerical input parameters to be initialized to only one of five values, ranging between 0 and 1 or -1 and 1. Part of this decision was to reduce the space and the other part was to make interpreting qualitative data more tractable by turning the numerical input parameters into ordinal parameters.